Microbiome Analysis Infrastructure Roadmap for Australia

V1.1

30 November 2020

Tiffanie M Nelson and Jeffrey H Christiansen

Doi: [to be added to the final version (V4.0)]

Contents

1. Executive Summary

Microorganisms are essential to life and they play important roles in many different environments. A ‘microbiome’ is an entire habitat, including the microorganisms (bacteria, archaea, eukaryotes, and viruses), their genomes (i.e. genes), and the surrounding environmental conditions.

Analysis of microbiomes requires an environmental sample to be collected (e.g. soil, water, on/in host organisms) followed by nucleic acid sequencing from the sample. In regards to sequencing, there are two broad methods used: targeted approaches (amplicon analysis) and random shotgun (metagenomics). The choice of sequencing approach impacts subsequent analysis options to determine the diversity, abundance and function of microorganisms within and between samples.

In Australia, microbiome analysis is increasingly conducted across a wide variety of environmental (e.g. soil, water, etc), host-associated (e.g. animal, plant, coral, etc), and clinical sources. This document includes:

- a brief summary of microbiome analysis tools and methodologies,
- how the Australian community currently undertakes this work and their common data-, software- or compute-related infrastructure challenges (information obtained through consultation with a ‘Special Interest Group’ (SIG) of researchers undertaking microbiome analyses across Australia), and
- a high-level description of key components of an envisaged shared national microbiome analysis infrastructure for Australia, which, when implemented, would enable Australian researchers from a wide range of institutions to perform microbiome analyses work they would otherwise be unable to undertake because of the reported data-, software- or compute-related roadblocks, i.e.

  D1. A platform for performing microbiome analyses: to provide all Australian researchers with access to a shared platform with tools and workflows for microbiome analyses, underpinned by sufficient compute resources and easily connectable to a variety of data storage locations and key datasets from public repositories.

  D2. Systems for statistical analysis and visualisation of microbial communities: to make it easier for Australian researchers to perform relevant statistical and/or visualisation-based analyses of microbiome / microbial community data.

  D3. Systems to enable submission of raw sequencing reads and metagenome-assembled genome files from Australia to appropriate global repositories: to make it easier for any Australian researcher to share and publish their metagenome and microbiome-related data files publicly in accordance with best-practice open science guidelines.

Feedback on the proposed components outlined in this initial draft plan is now sought from the SIG and any other Australian researchers undertaking microbiome analyses. Following engagement with other stakeholder groups (i.e. international entities operating microbiome analysis infrastructure elsewhere and Australian research IT infrastructure partners), further iterations of this document will be produced with a final version of the plan scheduled for February 2021.

2. Background and Context

In Australia, investments to establish community-scale infrastructure to support bioinformatics-based research have materialised in various forms and scales over the last decade under a range of funding schemes. One significant supporter is Bioplatforms Australia[1], which aims to develop and support Australia's national bioinformatics infrastructure and is funded under the National Collaborative Research Infrastructure Strategy (NCRIS)[2].

Since 2019, Bioplatforms Australia has supported the Australian BioCommons[3], which is an initiative focussed on establishing improved access to bioinformatics tools, methods, datasets, computational infrastructure, and training for Australia's molecular life scientists to underpin world-class science. The Australian BioCommons is currently coordinating several national consultations with various communities of practice to gain input from life science researchers, bioinformaticians, and infrastructure providers to identify, configure, connect and support infrastructure to support bioinformatics-based research and resources that are relevant to these research communities.

To support the large (and growing) community of practice in Australia undertaking microbiome analyses, in late July 2020, the Australian BioCommons convened a "Microbiome Analysis Special Interest Group (SIG)" and invited participation from over 100 researchers across Australia with either experience in, or interest in microbiome research[4].

The outcome of the survey and that meeting is this document, which summarises and represents the current or expected infrastructure roadblocks and challenges described by members of the community, and identifies the potential broad features and requirements for shared, national infrastructure solution options that could help address these challenges.

Community input is welcomed at all times, as is the nomination of additional members of the SIG, by either adding comments directly to this google document, or by emailing communities@biocommons.org.au

Feedback on the proposed components outlined in this initial draft plan is now sought from the SIG and any other Australian researchers or their collaborators undertaking metagenomics and microbiome analyses.

Following engagement with other stakeholder groups (i.e. international entities operating metagenomics and microbiome analysis infrastructure elsewhere and Australian research IT

[1] Bioplatforms Australia
[2] National Collaborative Infrastructure Strategy (NCRIS)
[3] Australian BioCommons
[4] see Section 3.2 for methodology employed for formation of the group and membership

infrastructure partners), further iterations of this document will be produced with a final version of the plan scheduled for February 2021.

3. Microbiome Analysis - Methods and Community

3.1 What is microbiome analysis, how is it done, and why?

A 'microbiome' is an entire habitat, including the microorganisms (bacteria, archaea, lower and higher eukaryotes, and viruses), their genomes (i.e., genes), and the surrounding environmental conditions[5].

Analysis of microbiomes requires an environmental sample to be collected (e.g. soil, water, on/in host organisms) along with appropriate contextual information about the experimental data, known as metadata or the who, what, when, where and why of these data[6]. The metadata exists at multiple stages along the path of a microbiome study indicating the process of the analysis. Initial sample collections from an environment include sample metadata (e.g. soil/water depth, temperature, host characteristics, etc), followed by preparation metadata (e.g. information about the DNA extraction or nucleic acid sequencing methods from the sample)[6]. Following sequencing, reads are then analysed using computational methods to determine the diversity, abundance, and function of microorganisms within the sample and between samples/environments. At this stage, data processing and feature metadata (e.g. software and tools and output data table parameters)[6] are recorded.

Due to challenges in cultivating many microorganisms *in vitro*, the application of this approach to directly assay environmental and host samples has dramatically enhanced understanding of many microbial communities[7].

The majority of microbiome projects to date have used short-read technology (up to 600 bases) where long-read technologies (up to 10,000 bases) are only being applied more recently as the technology develops. For the purpose of this Roadmap, our focus will be on the short-read technologies that are the current standard practice in the discipline. Of the short-read sequencing technologies, there are two broad methods used: targeted approaches (amplicon analysis) and random shotgun (metagenomics). The choice of sequencing approach impacts subsequent analysis approaches.

3.1.1. Targeted methods (amplicon analysis)

For the targeted sequencing approach, specific marker genes found in certain taxa are amplified and used for identification and classification of those taxa (and no others) in the sample, e.g. 16S ribosomal RNA genes for Bacteria and Archaea. Amplicon profiling as a phylogenetic marker of bacteria, archaea, and fungi has proven to be a cost-effective and computationally

[5] Marchesi, JR. & Ravel, J. 2015, doi.org/10.1186/s40168-015-0094-5
[6] National Microbiome Data Collaborative's Introduction to Metadata and Ontologies, microbiomedata.org/introduction-to-metadata-and-ontologies/
[7] Parks, DH. et al. 2017, nature.com/articles/s41564-017-0012-7

efficient strategy for microbiome analysis[8] and may also allow for the imputation of functional genes based on their taxon in human studies[910]. Pipelines and workflows for amplicon analysis (Figure 1) continue to evolve, despite their establishment more than two decades ago[11]. Choice of hypervariable regions of marker genes, sequencing technology platform, workflow pipeline, software package(s), and database choice for taxonomic classification can all impact amplicon-based microbiome analysis outputs with respect to reproducibility and accuracy. While amplicon profiling has been widely used for many years, there is a rapidly growing interest in the random shotgun sequencing approach because it yields far greater insight from a sample.

3.1.2. Shotgun sequencing (metagenomics)

For a shotgun sequencing approach, any part of any genome in the environmental sample is sequenced and the resulting broad range of sequences can provide rich information. With deep sequencing, there is the potential for assembling complete 'metagenome-assembled genomes' (MAGs) for the many species within the sample, which yields contextual functional gene information (e.g. full-length protein sequences, gene context, and identification of pathways or gene clusters that may span more than a single contig) and contributes to uncovering the diversity of microorganisms (inclusive of bacteria, archaea, eukaryotes such as fungi, and viruses)[12,13,14].

Shotgun metagenomics offers the advantage of species or strain-level classification with greater accuracy and allows the functional content of samples to be determined. However, it is comparatively more expensive (due to the increased level of sequencing required), and processing shotgun metagenomic data into understandable taxonomic and functional profiles requires far greater computational infrastructure than that required for targeted sequencing. For instance, shotgun experiments can yield hundreds of millions of sequences with tens of gigabytes of data[15], and methods used to generate MAGs (Figure 1) can require gigabytes to terabytes of computational memory[16] to assemble genomes using software such as, metaSPAdes[17]. However, less memory intensive solutions, such as MEGAHIT[18], MetaVelvet-SL [19] and IDBA-UD[20] are continually made available and developed by the bioinformatics community. As the cost-efficiency of next generation sequencing continues to improve adn tools and software are continually accessible, metagenomic (shotgun) sequencing is increasignly becomignt he method of choice[21]

[8] Rausch, P. et al. 2019, microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0743-1
[9] Sun, S. et al. 2020, microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00815-y
[10] Douglas, GM. et al. 2020, nature.com/articles/s41587-020-0548-6
[11] Schloss, PD. et al. 2009, aem.asm.org/content/75/23/7537
[12] Thomas, T. et al. 2012, ncbi.nlm.nih.gov/pmc/articles/PMC3351745/
[13] Olm, M. et al. 2019, microbiomejournal.biomedcentral.com/articles/10.1186/s40168-019-0638-1
[14] Schulz, F. et al. 2020 www.nature.com/articles/s41586-020-1957-x
[15] Mitchell, AL. et al. 2017, academic.oup.com/nar/article/46/D1/D726/4561650
[16] Mitchell, AL. et al. 2019, academic.oup.com/nar/article/48/D1/D570/5614179
[17] Nurk, S. et al. 2017, ncbi.nlm.nih.gov/pmc/articles/PMC5411777/
[18] Li, D. et al. 2015, academic.oup.com/bioinformatics/article/31/10/1674/177884
[19] Afiahayati, et al. 2015, pubmed.ncbi.nlm.nih.gov/25431440/
[20] Peng, Y. et al. 2012, academic.oup.com/bioinformatics/article/28/11/1420/266973
[21] Vollmers, J. et al. 2017, journals.plos.org/plosone/article?id=10.1371/journal.pone.0169662

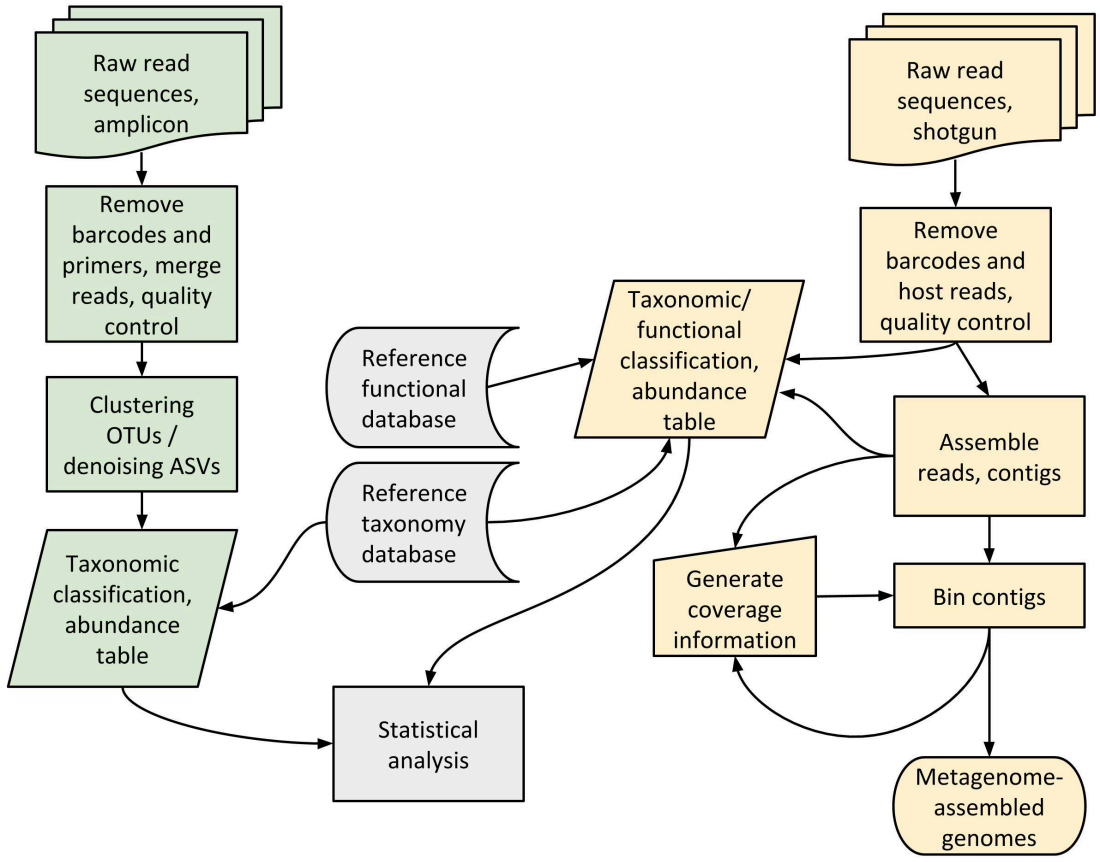

Figure 1: Typical amplicon (targeted) and metagenomic (shotgun) workflows

The amplicon (targeted) workflow is shown on the left in green and the metagenomic (shotgun) workflow on the right in yellow. Components shown in the centre in grey are common to both workflows. The workflows display the dominant steps used to transfer raw sequence reads from marker gene surveys or shotgun metagenomic sequencing into abundance tables for taxonomic classification, functional profiling, or metagenome-assembled genomes. Access to numerous tools supported by computational infrastructure may be employed at each of the various steps, with information feeding back into previous steps more than once. Created in part from information detailed in Liu, Y-X. et al. 2020[22]

3.2 Who in Australia is performing microbiome analysis, and which species are they tackling?

The ability to extract DNA directly from environmental samples, apply high throughput shotgun or amplicon sequencing and conduct subsequent microbiome analyses has greatly advanced our knowledge of the micro-community including archaea, bacteria, fungi, and viruses inhabiting various environments. The benefits of this approach have a far-reaching impact on the

[22] Liu, X-Y. et al. 2020, link.springer.com/article/10.1007/s13238-020-00724-8

environment, agriculture, and health, with the suggestion that the promise of many benefits is still to come[23].

Broad benefits include a greater understanding of the genetic identity and phylogeny of microorganisms in a sample allowing for an elucidation of the impact or relationship to the host or source environment[24]. A few specific examples include direct clinical diagnosis of an infection, e.g. efficient detection of meningoencephalitis by rare bacteria[25]; transmission network analysis to investigate disease outbreaks, e.g. source tracking to prevent further transmission of a deadly pathogen locally[26] or globally during the coronavirus pandemic[27]; “nature mining” or bioprospecting for biologically active secondary metabolites for use in health remedies through the identification of bioactive enzymes to improve industrial processes, e.g. novel enzymes from seawater for use in the dairy industry[28]; and, identification of novel bio-indicator species to target resources for environmental conservation[29].

Hence, the critical importance of developing the molecular techniques to identify and study microorganisms through their genetic material is a key methodology to help to address challenges of strategic importance to Australia, and as such is touched on in several Australian Academy of Science Decadal Plans for Science[30]: Biodiversity[31], Agricultural Science[32], Marine Science[33], Ecoscience[34], Nutrition Science[35] and Geoscience[36]. Assembling metagenomes in whole or in part from a wide and diverse range of organisms will be a key process that must be undertaken to fully realise the application of microbiome analysis within this vision.

The advent of affordable sequencing is enabling microbiome analysis to be applied as a routine method for groups working on a variety of environments and host organisms. Many groups and consortia across Australia are now actively working on producing high-quality microbiome and MAG datasets, with a general focus on Australian ecosystems, particularly soils and marine

[23] Chiu, CY. and Miller, S.A. 2019, nature.com/articles/s41576-019-0113-7
[24] Simon, C. and Daniel, R. 2011, aem.asm.org/content/77/4/1153
[25] Wilson, MR. et al., 2014, nejm.org/doi/full/10.1056/NEJMoa1401268
[26] Loman, NJ. et al. 2013, jamanetwork.com/journals/jama/article-abstract/1677374
[27] Rockett, RJ. et al. 2020, nature.com/articles/s41591-020-1000-7
[28] Wierzbicka-Wos, A. et al. 2013, bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-13-22
[29] Kloet, R. 2010, tandfonline.com/doi/full/10.1080/1943815X.2010.542165#
[30] 10-year strategic plans for science disciplines, developed by the Australian Academy of Science’s National Committees for Science.
[31] science.org.au/support/analysis/decadal-plans-science/discovering-biodiversity-decadal-plan-taxonomy
[32] science.org.au/support/analysis/decadal-plans-science/decadal-plan-agricultural-sciences-2017-2026
[33] science.org.au/support/analysis/reports/national-marine-science-plan
[34] science.org.au/support/analysis/reports/foundations-future-long-term-plan-australian-ecosystem-science
[35] science.org.au/supporting-science/science-policy-and-analysis/decadal-plans-science/nourishing-australia-decadal-plan
[36] science.org.au/supporting-science/science-policy-and-sector-analysis/decadal-plans-science/australian-geoscience

systems[37,38], but also human health[39,40,41,42,43,44,45] as well as programs that study microorganisms in any habitat[46,47,48].

The scientific literature indicates an approximate number of studies using shotgun metagenomics or amplicon/marker gene surveys produced from Australian-based researchers (see Figure 2).

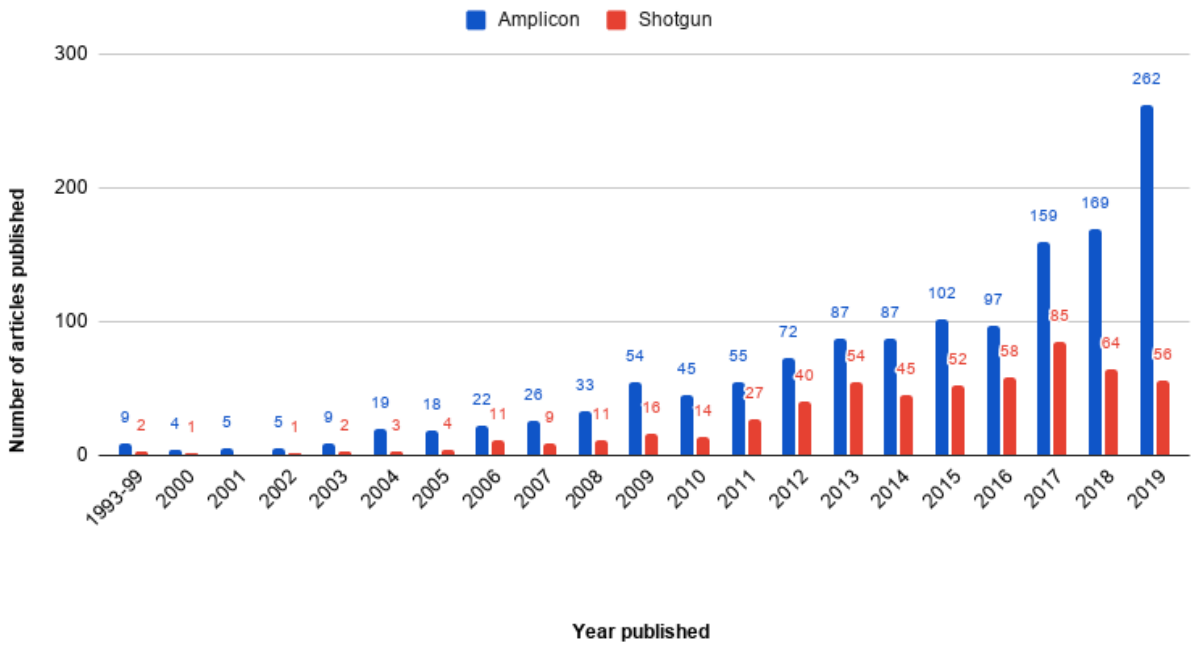

Figure 2: Estimates of the increasing number of microbiome analysis studies conducted in Australia

To gain an estimate of the number of microbiome analysis studies that have been conducted historically in Australia, a search was conducted of the Scopus database for articles with: either A/ 'shotgun' or 'metagenomic' in the title, abstract, or keyword and 'Australia' in the affiliation; or B/ 'amplicon' or 'microbiome' or 'microbiota' or 'microbial community' or 'virome' in the title, abstract or keyword and genome or sequencing or sequence or genomic or next-generation in the title, abstract or keyword and 'Australia' in the affiliation. Articles retrieved from the search were manually reviewed to include only those whose focus included the production of data using a marker gene or metagenomic sequencing method and excluded others whose focus was on developing or evaluating analysis methods or tools. Articles that were retrieved during multiple searches were limited to include only one representative article categorised to either marker gene or shotgun. The complete list of citations including abstracts can be found here.

[37] Australian Microbiome Initiative
[38] Marine Microbiome Initiative
[39] Microbial Biology and Metagenomics, Translational Research Institute
[40] Microbiome Research Centre
[41] Marshall Centre, University of Western Australiabeneficial-microbes
[42] Aussie Gut
[43] The Healthy Optimal Australian Microbiome (HOAM) Study
[44] Charles Perkins Centre
[45] Microbiome and Metabolome of Pregnancy and Early Life
[46] ithree institute
[47] Australian Centre for Ecogenomics
[48] Centre for Microbiome Research

In late July 2020, the Australian BioCommons invited over 100 researchers across Australia to participate in a Microbiome Analysis Special Interest Group (SIG). These researchers were identified as having experience in, or interest in, microbiome analysis. The Australian BioCommons sought information from the SIG about each member's level of expertise, current (and desired) practices and infrastructure used via an on-line survey[49] (number of respondents = 33), and also held an open video conference follow-up to gain further information (minutes[50] and a recording[51] of the meeting are available).

Respondents to the survey and attendees at the meeting collectively indicated they are performing microbiome analyses on both samples from environmental (i.e. marine, freshwater, soil, and air) as well as host-associated (e.g. animals, plants, corals and humans) habitats. The collective responses also indicated that all of the following approaches are being undertaken by Australian researchers: targeted amplicon sequencing, random shotgun sequencing, taxonomic profiling, functional profiling, generating metagenome-assembled genomes (MAGs), phylogenetic analysis, statistical analyses, and novel gene discovery.

3.3 How is microbiome analysis being done in Australia?

3.3.1 Data

Based on information received from the SIG members through the survey (*n*=33), most researchers use a combination of sequencing platforms to generate their data with the most popular being Illumina[52], Nanopore[53], and PacBio[54].

Researchers depend on access to up-to-date databases that house information for classifying gene sequences to identify taxonomy or function, and collectively, the SIG identified accessing more than 36 different databases[55].

To aid in taxonomic classification, 82% of respondents indicated the use of the National Center for Biotechnology Information (NCBI) database of raw sequences: the Sequence Read Archive[56] (SRA).

[49] Presentation including survey results on microbiome analysis infrastructure needs and challenges conducted 15/06/2020 to 31/07/2020 is here.
[50] Meeting minutes from microbiome analysis SIG meeting held 23/07/2020 are here.
[51] Recording of microbiome analysis SIG meeting held 23/07/2020 is here.
[52] Illumina, Inc.
[53] Oxford Nanopore Technologies
[54] PacBio, Pacific Biosciences
[55] Complete list of databases used by survey respondents with number of responses shown in brackets: NCBI (27), KEGG (24), Silva (18), Pfam (18), RDP (13), COG (13), KOG (13), TIGRFAMs (12), Greengenes (12), EBI (12), Custom-made databases (12), SEED (8), eggNOG (6), GTDB (5), PR2 (2), MetaCyc (2), CAzy (2), TCDB (1), TARA Oceans (1), Rfam (1), MEROPS (1), MAR (1), FunGuild (1), databases integrated with InterPro (1), and Cyanorak (1).
[56] SRA, Sequence Read Archive

For functional classification, the Kyoto Encyclopedia of Genes and Genomes[57] (KEGG) databases (which provides information relating to the functional classification of cells and organisms), is accessed by 72% of survey respondents.

3.3.2 Tools

Based on the survey, approximately 100 software tools, pipelines, or packages were identified as being used by respondents for various stages[58] of the microbiome analysis process. These are listed in Appendix 1 of this document.

The data generated in either amplicon marker gene or shotgun metagenomic surveys present a wide variety of possible analysis pathways/workflows to pursue and there are many options for tools/pipelines or processes at each step of a chosen bioinformatic pathway.

In the early part of these workflows, the choice of a specific tool is often dictated by the sequencing platform/s that was/were used for data generation. Some respondents (40%, n=13) noted that custom tools developed within their group were sometimes necessary for the latter stages of their workflows due to the highly novel nature of the metagenomes being studied and the lack of available tools for shotgun datasets, especially when taking a systems biology approach to the research.

A number of software packages (e.g. QIIME2[59], Mothur[60]) or web-based platforms (e.g. MG-RAST[61]) that include numerous 'wrapped' tools were also identified by the SIG as complete processing or automatable workflows that convert raw sequences to output abundance tables of taxonomic or functional classifications, visualisations or associated data products.

Several researchers (36%, n=12) reported that they were not using their preferred tools/pipelines (primarily due to not having access to sufficient computational memory to run these tools - see Section 3.4.1) and instead had resorted to a workaround solution with other tools.

3.3.3 Compute infrastructure

3.3.3.1 Types used

Survey respondents (*n* = 33) currently use a variety of computational infrastructure for their analyses. Most access high-performance computing provided by their host institute (85%) or use their lab laptops or personal computers (73%), with fewer respondents (36%) using shared high-performance computers managed by national (e.g. NCI[62], Pawsey[63]) or state (e.g.

[57] KEGG: Kyoto Encyclopedia of Genes and Genomes
[58] e.g. quality control, preprocessing, OTU/ASV picking clustering, taxonomic classification, sequence assembly, gene prediction and alignment, annotation prediction, assembly validation, statistical analysis and visualisation
[59] QIIME2.0
[60] Mothur
[61] MG-RAST
[62] NCI, National Computational Infrastructure, nci.org.au/
[63] Pawsey Supercomputing Centre, pawsey.org.au/

QCIF/QRIScloud[64]) computing centres or accessing commercial cloud resources (21%), such as Amazon Web Services (AWS)[65]. All respondents (100%) use more than one of these compute-infrastructure types to support their work and mix and match their use to the problem at hand.

3.3.3.2. Resourcing

More than half of the respondents (62%, *n* = 16) said the infrastructure they currently had access to was not sufficient for their current *metagenomic* work, due to limitations in available memory, data storage allocations, or being able to access relevant databases such as the SRA in a workable manner for locally based computing.

3.4 Challenges being faced

A variety of limitations/roadblocks/challenges/issues with current infrastructure were identified by the SIG.

3.4.1 Computational resourcing and set-up challenges

- Computational resources available (even across a variety of infrastructures) can be insufficient, especially when processing MAGs/complex communities which require many CPUs and RAM (e.g. up to 3TB RAM) to allow for the co-assembly and alignment of numerous sequences with samples/datasets. Workarounds include either limiting the dataset size or accessing commercial Google Cloud Platform (GCP)[66] and Amazon Web Services (AWS) clouds which incurs a cost with each analysis;
- Some respondents perceive that resource allocation practices undertaken by HPC providers lead to poor utilisation equality among users, and that "*fair and transparent user resource allocation*" and "*intelligent and active resource management*" was lacking;
- Some respondents also perceive a lack of expertise in the build, maintenance, and management of some computational resource providers for metagenomics analyses and this creates bottlenecks and challenges for troubleshooting;
- Obtaining access to computational resources through Tier 1 (i.e. NCI, Pawsey) resource infrastructure for metagenomics projects can present difficulties due to challenges in establishing benchmark metrics on the anticipated resources;
- Ethics applications for intended metagenomic or microbiome analyses in human research studies require clear data management and security protocol information from computational infrastructure providers, yet this information is not readily available from the providers.

[64] Queensland Cyberinfrastructure Foundation, QCIF
[65] Amazon Web Services
[66] Google Cloud

3.4.2 Data related challenges

- Accessing raw read data for micro-organisms housed in the (USA-based) large and continuously growing SRA database[67] to analyse or mine on locally available computational infrastructure is limiting due to slow data movement for transfer of very large volumes of data from the USA to Australia. Some SIG members noted that SRA data is also available in the cloud and can be accessed via the commercial Google Cloud Platform (GCP) and Amazon Web Services (AWS) clouds, which require a virtual machine instance to be set up, and payment for the use of these commercial services. Several researchers indicated that more efficient and/or cost-effective access to the SRA in Australia would increase research output by providing opportunities for new projects, such as whole data mining or available sequences.

- Data publishing from Australia to international repositories (e.g. GenBank/SRA) is considered by some to be difficult - primarily due to an unclear submission process, changing input requirements, and issues with uploading the data to the repositories.

- Other data related challenges include inefficient access to databases for classification purposes (*n* = 6); a complete lack of relevant databases for taxonomic identification of some groups of organisms (*n* = 3), and a lack of metadata (either collected or fully recorded) to enable appropriate data reuse (*n* = 1).

3.4.3. Other challenges

- Tools to enable better data or methods management are generally lacking, with more than two-thirds of the respondents (*n* = 12) reporting that no specific data or method management tools or frameworks are used to support their microbiome analysis projects. From the researchers who responded to the survey question on this topic, (*n* = 2), Jupyter Notebooks[68], Bitbucket[69] or GitHub[70] or R[71] were used for methods/code management.

- Other frustrations raised by the SIG include that many tools are too human-centric and do not provide enough information for non-model, non-human environments (*n* = 3), or that investment is lacking to enable the development of more efficient metagenome assembly algorithms that don't require so much RAM (*n* = 1).

[67] Sequence Read Archive at National Center for Biotechnology Information
[68] Jupyter Notebooks
[69] Bitbucket
[70] GitHub
[71] R Statistical Project

3.5 Is a shared national solution palatable to the research community?

All but two of the respondents (93%, *n* = 27) agreed that if a shared data collaboration/analysis platform for microbiome analysis was available for use, they would use such a platform provided it was well designed and supported. This number included respondents who stated that their needs are currently met.

Twenty-two hypothetical features of such a shared system are listed in Figure 3, ranked according to how crucial respondents believe that feature would be (when asked would the feature be 'crucial', 'important' or 'unimportant'). The top several features of a shared platform deemed the most crucial are: (1) ease of uploading/downloading data, (2) ease of access from anywhere; (3) an ability to transfer data to/from storage; (3) access to preferred tools/pipelines as well as (4) quick installation of tools/pipelines on request; (5) good documentation on platform use; (6) long term support for the platform; and (7) easy self-management of permissions/access to collaborators.

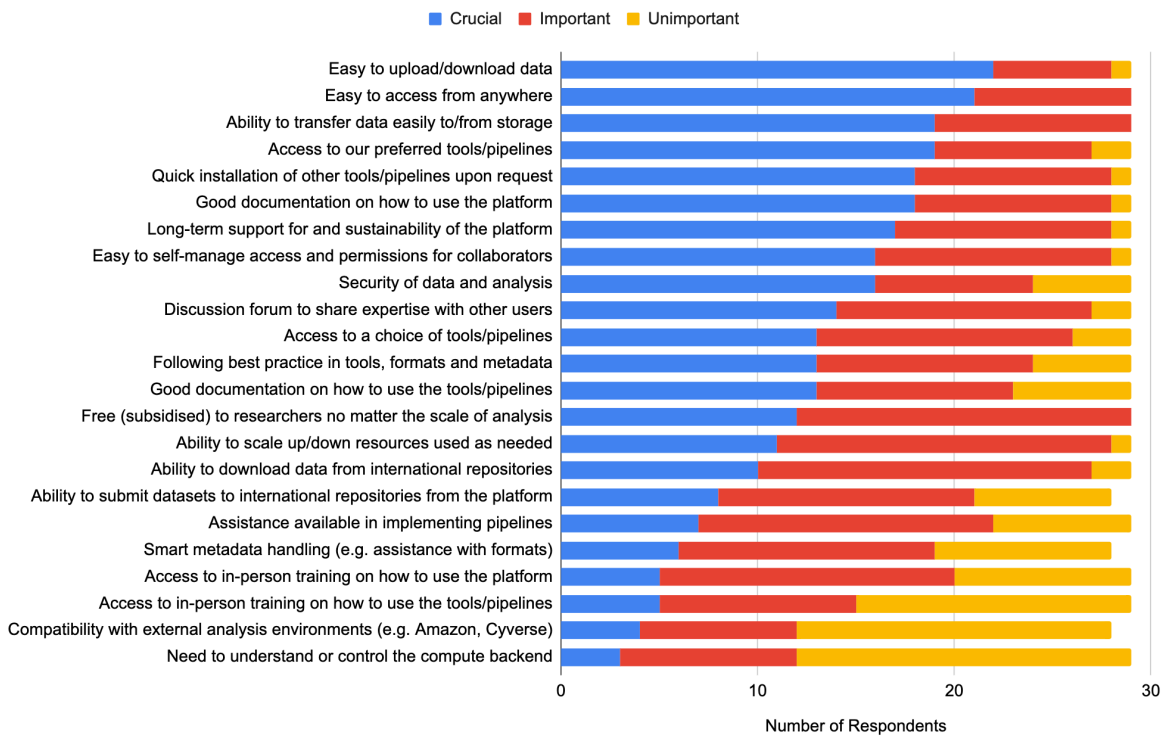

Figure 3: Desired features in a shared metagenomic/microbiome analysis platform

Survey respondents were asked about which features they considered to be 'crucial', 'important' or 'unimportant' in a shared microbiome analysis workspace. The number of responses classified at each level is shown per feature, and features are ranked in descending order from those deemed to be most crucial to the least crucial.

4. Meeting the Needs of Australian Researchers for High-quality, Accessible Microbiome Analysis Infrastructure

4.1 Goal

The Australian BioCommons aims to develop a ‘Microbiome Analysis Infrastructure Roadmap for Australia’ that describes collaborative infrastructure, which, when implemented (from Q1 2021 onwards), will enable Australian researchers from a wide range of institutions to perform high-quality microbiome analysis work who would otherwise be unable to do so because of data-, expertise-, software- or compute-related infrastructure roadblocks.

Four versions of the Roadmap document are planned, each to incorporate content and feedback from different groups. Planned dates for the development of the Roadmap are as follows:

- V1 (this document) - Content-based on SIG survey results and input from SIG meeting - November 2020.
- V2 - Content modified to incorporate feedback from SIG, other researchers undertaking metagenomics and microbiome analysis, and international groups - December 2020/January 2021.
- V3 - Content modified to incorporate feedback from various national computational infrastructure providers - December 2020/January 2021.
- V4 - Content modified to incorporate final feedback from SIG - February 2021.

4.2 Objectives

The high-level objectives of deploying the proposed infrastructure and associated services are:

1. To provide Australian researchers with access to a platform with:
    a. A selection of tools and workflows that will allow microbiome analyses (whether they be amplicon/targeted or shotgun/metagenomics based) to be performed across a wide range of taxa;
    b. Sufficient computational infrastructure and resources; and,
    c. Connectivity to a variety of data storage locations (locally and internationally).
2. To make it easier for Australian researchers to perform statistical and visualisation analyses of microbiome data; and,
3. To make it easier to publish high-quality microbiome-associated data files in accordance with best-practice open science guidelines.

4.3 Outputs

To address the objectives, three broad outputs/infrastructure components are proposed for implementation:

D1. A platform for performing taxonomic and functional analyses of microbiomes

D2. Systems to enable statistical analyses and visualisation of microbial community data

D3. Systems to enable submission of raw sequencing reads and metagenomic assembled genome files from Australia to appropriate global repositories

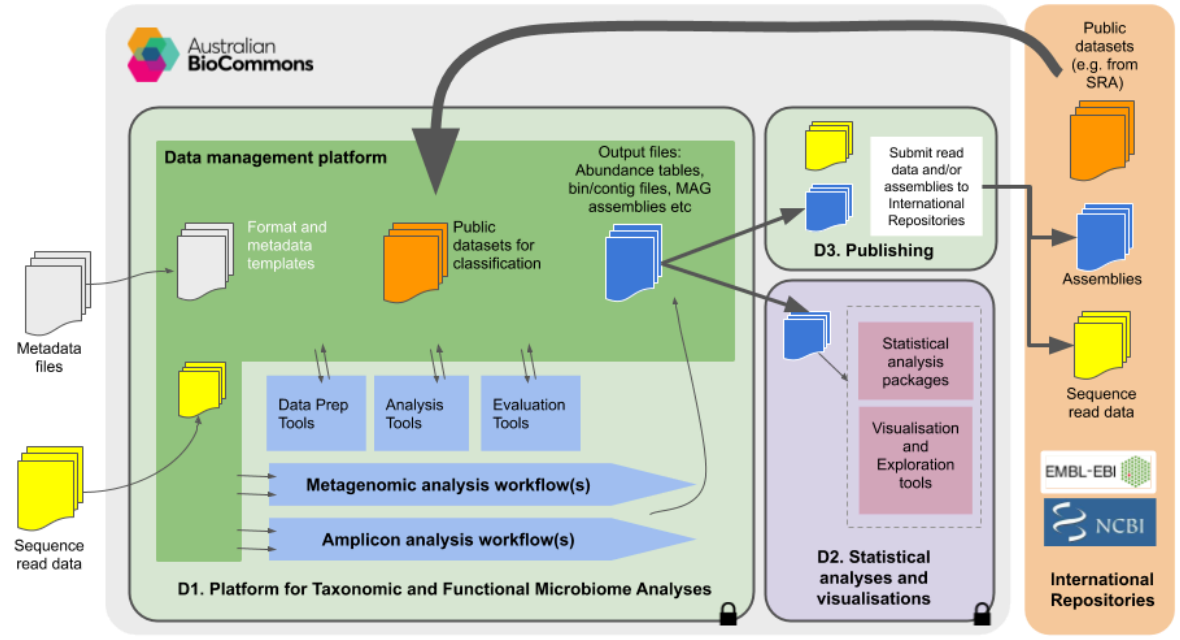

Figure 4. Schematic diagram showing the proposed infrastructure to support microbiome analyses, and data flow

(D1) Sequence reads or other relevant data are inputs into the Platform for Taxonomic and Functional Microbiome Analyses which provides a command-line interface (CLI)- or graphical user interface (GUI)-based access to tools and workflows for performing amplicon marker gene clustering or metagenome assembly and classification (blue shapes). It is underpinned by sufficient and appropriate computational infrastructure. Closely associated is a data management platform (denoted by the darker green shape) that caters to data management, version control, and association of appropriate (e.g. sample, experimental) metadata with the data files. Outputs of D1 are accessible to both (D2) hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of microbiome datasets, and (D3) systems to enable submission/publishing of metagenome-assembled genome files (and sequence read data) to international repositories. Arrows indicate the general flow of data. Thicker arrows indicate increasing data transfer capabilities. See Appendix 1 for a list of tools/pipelines that may be included in D1. Higher resolution image.

D1 - A platform for performing taxonomic and functional analyses of microbiomes;

To address objective 1 (*i.e. providing Australian researchers with access to a selection of tools and workflows underpinned by computational resources that allow taxonomic and functional analyses of microbiomes (whether they be derived from amplicon/targeted or shotgun/metagenomics based sequencing approaches) to be performed*), it is proposed to implement a platform in Australia[72], that:

[72] Subject to the results of a platform functionality comparison/gap analysis, scoping of compute requirements, agreement with various computational providers about hosting, and outcomes of further consultation with end users.

A. Includes a set of key tools[73] and/or pipelines[74] for data preparation, quality control, metagenome/microbiome functional or taxonomic table abundance table production, classification, and production of metagenome-assembled genomes:

    a. Installed (plus all other dependencies) and optimised on a command line interface (CLI) analysis environment (i.e. across a variety of Tier 1 and 2[75] shared computational infrastructures) underpinned by appropriate computational resources[76];

    b. Installed (plus all other dependencies) and optimised on a graphical user interface (GUI) web-based data analysis platform where possible, (i.e. Galaxy Australia[77]), underpinned by appropriate computational resources; and,

    c. Available as high quality, trusted software containers for self-deployment on institutional or independent computational infrastructures.

B. Has support available from experts for installation/containerisation of extra software tools and maintenance with version control and updates as required;

C. Is easily connectable to a variety of data storage locations, including public databases, i.e. international (e.g. SRA, either at NCBI or in the cloud)[78], national (i.e. CloudStor[79]), institutional or other data storage, and with the ability to upload/mount user-generated or other datasets[80] that are required as inputs for a microbiome analysis pipeline;

D. Has appropriate user authorisation and sharing mechanisms to allow for data sharing, solely at the discretion of a data owner/custodian;

E. Is tightly associated with a data management component that contains shared metadata templates that include all elements required to enable submission of files to international repositories, when required;

F. Support available from experts in formatting data and curating metadata to comply with NCBI/ENA repository format requirements[81];

---

[73] e.g. a selection of the tools listed in Appendix 1
[74] e.g. MGnify pipelines
[75] The definition of Peak (Tier 1) High Performance Computing (HPC) is traditionally defined as a compute capability that is in the top 200 globally. Australia's current Tier 1 facilities are: NCI and Pawsey. Examples of Tier 2 facilities include State-level systems such as QRIScloud operated by QCIF and many institutionally operated facilities.
[76] Including necessary high memory nodes (>1TB RAM) for performing large coassembly research and phylogenetic tree creation. See also biocommons.org.au/pathfinder-biocloud
[77] Galaxy Australia
[78] Note that previous work commissioned by Queensland Genomics established a pilot implementation of a local Australian SRA Cache - see Cuddihy T. *et al* (2019) Stud Health Technol Inform doi:10.3233/SHTI190776.
[79] CloudStor Service
[80] Including necessary datasets for classification, comparison of genome sequences or unique data mining research projects.
[81] potentially building on the previous data submission service which was offered nationally by the EMBL-ABR: QCIF node, and is now available to researchers from QCIF/QFAB member organisations

G. Includes documentation, including a knowledgebase with community-contributed content; and,

H. Includes training for all the above.

D2. Systems to enable statistical analyses and visualisations of microbial community data:

To address objective 2 (*i.e. to make it easier for Australian researchers to perform statistical and visualisation analyses of microbiome data*), it is proposed to implement:

A. Hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of microbiome datasets[82];

B. Appropriate user authorisation and sharing mechanisms to allow for public or private data and associated data product(s) sharing, solely at the discretion of a data owner/custodian;

C. Documentation on how to use the system (including a knowledgebase with community-contributed content); and,

D. Training.

D3 - Systems to enable submission of raw sequencing reads and metagenome-assembled genome files from Australia to appropriate global repositories:

To address objective 3 (*i.e. to make it easier to publish high quality and share final raw metagenome-assembled genomes (and relevant input data) in accordance with best-practice open science guidelines*) it is proposed to implement:

A. A temporary 'staging post' in Australia for metagenome and microbiome (and sequence read) files ready for public international release. The system should include data/metadata formatting checks (which would be enabled by the use of the data management platforms described in D1-E), and support as detailed in D1-F;

B. Includes a rapid data transfer from the data management platform or the sharing platform to NCBI and/or ENA; and,

C. Documentation on how to use the system (including a knowledgebase with community-contributed content).

[82] e.g. packages written in R (e.g. phyloseq) or python, or systems such as Phinch, Anvi'o etc

4.4 Implementation timeframes

It is intended that the components identified in Section 4.3 will be implemented throughout 2021-2022.

As of November 2020, several key activities that are relevant to the proposed infrastructure are already underway:

Component	Planned dates for delivery	Notes
D1-Aa. Key tools/workflows installed as modules and optimised for CLI access across a variety of Tier 1 and Tier 2 HPC infrastructures.	Ongoing	As of November 2020, 6 of the tools listed in Appendix 1 (graftm, groopm, metacv, QIIME, QIIME2.0, SortMeRna) are installed as modules on QRIScloud/UQ-RCC HPC machines (Tinaroo[83], Awoonga[84], FlashLite[85]). Installation of further tools as modules across NCI, Pawsey, and QRIScloud/UQ-RCC infrastructures to support microbiome analysis is being undertaken in the BioCommons 'BYOD' Expansion Project. Preliminary discussions have been held with the MGnify group at EBI[86] to install and host a MGnify 5.0 pipeline (which offers specialised workflows for three different data types: amplicon, raw metagenomic/ metatranscriptomic reads, and assembly) on Australian BioCommons associated infrastructure, as well as the Marine Metagenomics group from ELIXIR-Norway surrounding the local installation of the Meta-Pipe workflow (for pre-processing, assembly, taxonomic classification and functional analysis of marine metagenomics data).
D1-Aa. CLI platform appropriately resourced for performing microbiome analyses	Ongoing	BioCommons partner infrastructures at NCI, Pawsey, and QCIF include machines that are capable of performing any part of microbiome

[83] Tinaroo high performance computer.
[84] Awoonga high performance computer.
[85] FlashLite high performance computer.
[86] European Bioinformatics Institute, EBI, is part of the European Molecular Biology Laboratory, EMBL, and is sometimes referred to as EMBL-EBI.

		analysis. This includes FlashLite at QCIF/UQ which can be structured to allow 'supernodes' of up to 8TB) Enabling increased access to partner HPC systems via mechanisms other than through the National Computational Merit Allocation Scheme (NCMAS) or partner shares are under active exploration by the BioCommons.
D1-Ab. Key tools/workflows installed as modules and optimised on Galaxy Australia.	Ongoing	As of November 2020, 4 of the tools listed in Appendix 1 (maxbin2, metaSPAdes, mothur, SortMeRna) are installed on Galaxy Australia. As of February 2021, a numerous of the tools listed in Appendix 1 (megahit, metaspades, qiime2 are installed on national infrastructure which can be viewed through the BioCommons tool Registry. Installation of further tools on Galaxy Australia can be requested by any member of the community at any time.
D1-Ab. Galaxy Australia appropriately resourced for performing microbiome analyses	Q1 2021	In addition to the 465 cores at QCIF, UMelb, and Pawsey that currently underpins Galaxy Australia, the Australian BioCommons has secured ARDC funding to purchase an additional minimum of 1x 4TB and 3x 2TB high memory nodes to contribute computational resources to Galaxy Australia. These nodes will be reserved for specific tools requiring high memory, such as those required for MAG assembly.
D1-Ac. Key tools available as high quality trusted software containers for self-deployment on institutional or independent computational infrastructures	Ongoing	Development of containerised tools to support various life science researcher communities in Australia (including microbiome analysis) is being undertaken in the BioCommons 'BYOD' Expansion Project.
D1-B. Connectable to Nationally available storage (e.g. Cloudstor)	Ongoing	In late 2020, a direct connection between Cloudstor and Galaxy Australia was implemented. Streamlined connectivity of Cloudstor storage to Pawsey, QCIF, NCI, and other computational resources will continue in the BioCommons 'BYOD' Expansion Project.

D1-C/D2-B. Appropriate user authorisation and sharing mechanisms	Ongoing	AAF is currently engaged by the BioCommons to explore Access and Authentication Frameworks that will be fit for purpose across all envisaged BioCommons-related platforms and services.
D1-G. Tool and software workflow documentation with community contributed content.	Ongoing	Tool and workflow documentation for other researcher communities (e.g. *de novo* genome assembly, and genome annotation) are being organised via an Australian BioCommons Github: https://github.com/australianbiocommons. This avenue is available for the microbiome analysis community.
D1-H. Training re. containerisation of software tools.	Ongoing	Introductory level training around software containerisation (co-organised by BioCommons and Pawsey) occurred in June/July 2020 and will be repeated throughout 2021, 2022, and 2023. See https://www.biocommons.org.au/events/containers-intro and the Australian BioCommons YouTube channel for recordings of these events.

As of November 2020, the following key activities are under active planning:

Component	**Notes**
D1-D. A data management system that is tightly linked to the Microbiome Platforms	Considerations for what may be the best technical solution are ongoing. See Requirements of a Data Management Component of the Australian BioCommons
D1-H Training re. taxonomic and functional bioinformatics of shotgun and targeted sequencing projects	Discussions with EBI to potentially deliver microbiome analysis related bioinformatics training events[87] to an Australian audience during 2021 or 2022 have begun.
D2-A. Hosted frameworks to enable researchers to utilise common packages for statistical analysis, visualisation, and exploration of microbiome datasets	'Interactive environments' offered through the Galaxy platform include R-Studio, JupyterLab, CloudStor SWAN[88], and Phinch. These are currently available publicly through the European public Galaxy instance (see https://live.usegalaxy.eu/), and are planned for release via Galaxy Australia in Q1 2021. Galaxy Interactive environments may represent an option for this feature.

[87] EBI has held the training courses 'Metagenomics Bioinformatics' and 'Introduction to Metagenomics' both virtually and also physically in multiple locations including Argentina in association with CABANA (Capacity Building for Bioinformatics in Latin America).
[88] CloudStor SWAN

D3-A and D3-B. A temporary ‘staging post’ in Australia for metagenome and microbiome (and sequence read) files ready for public international release, with a rapid data transfer from the data management platform or the sharing platform to NCBI and/or ENA	COPO is a GUI-based metadata platform for brokering life science data submissions to various repositories including the ENA (see https://f1000research.com/articles/9-495). It is being adopted by the Darwin Tree of Life project in the UK as the tool to enable the data and metadata submission to ENA to be completed for genome assemblies of over 60,000 species native to the British Isles. The Australian Biocommons is currently exploring whether a locally supported COPO instance can fulfill the requirements of D3-A/D3-B.

Appendix 1

Table 1. Microbiome analysis tools for consideration for inclusion in a shared analysis environment.

Note that a microbiome analysis protocol may also incorporate many other software tools not listed here.

Workflow Step	High-level component	Tool	Brief description	Link to data/software or article
1	Quality Control	FastQC	Provides a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines.	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2	Preprocessing	BLAST+	A suite of command line tools to run BLAST which is to search for nucleotide similarities.	https://blast.ncbi.nlm.nih.gov/Blast.cgi
2	Preprocessing	ChimeraSlayer	A chimeric sequence detection utility, compatible with near-full length Sanger sequences and shorter 454-FLX sequences (~500 bp).	http://microbiomeutil.sourceforge.net/
2	Preprocessing	fastp	Tool designed to provide fast all-in-one preprocessing for FastQ files.	https://github.com/OpenGene/fastp
2	Preprocessing	FASTX-Toolkit	A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.	http://hannonlab.cshl.edu/fastx_toolkit/
2	Preprocessing	FLASH - Fast Length Adjustment of SHort reads	A very fast and accurate software tool to merge paired-end reads from next-generation sequencing experiments.	https://ccb.jhu.edu/software/FLASH/
2	Preprocessing	MultiQC	A reporting tool that parses summary statistics from results and log files generated by other bioinformatics tools.	https://multiqc.info/docs/
2	Preprocessing	PANDAseq	A program to align Illumina reads, optionally with PCR primers embedded in the sequence, and reconstruct an overlapping sequence.	https://github.com/neufeld/pandaseq
2	Preprocessing	PEAR - Paired-End reAd mergeR	A fast and accurate Illumina Paired-End reAd mergeR.	https://cme.h-its.org/exelixis/web/software/pear/doc.html https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3933873/

2	Preprocessing	Prinseq	Easy and rapid quality control and data preprocessing of genomic and metagenomic datasets.	http://prinseq.sourceforge.net/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3051327/
2	Preprocessing	Prinseq++	A program to filter, reformat or trim genomic and metagenomic sequence data.	https://github.com/Adrian-Cantu/PRINSEQ-plus-plus
2	Preprocessing	SortMeRNA	A program tool for filtering, mapping, and OTU-picking NGS reads in metatranscriptomic and metagenomic data.	https://github.com/biocore/sortmerna
2	Preprocessing	Tagcleaner	A tool to automatically detect and efficiently remove tag sequences.	http://tagcleaner.sourceforge.net/
2	Preprocessing	Trimmomatic	A flexible read trimming tool for Illumina NGS data.	http://www.usadellab.org/cms/?page=trimmomatic
2	Preprocessing	UCHIME/ UCHIME2	Chimera detection tool.	https://www.drive5.com/usearch/manual/uchime2_algo.html https://www.biorxiv.org/content/10.1101/074252v1.full
2	Preprocessing	VSEARCH	Processes and prepares metagenomics, genomics, and population genomics nucleotide sequence data.	https://github.com/torognes/vsearch
3	OTU/ASV picking clustering	UPARSE	A method for generating clusters (OTUs) from next-generation sequencing reads	http://drive5.com/uparse/
3	OTU/ASV picking clustering	USEARCH	A unique sequence analysis tool with thousands of users worldwide.	https://www.drive5.com/usearch/
4	Taxonomic classification	Centrifuge	A very rapid and memory-efficient system for the classification of DNA sequences from microbial samples.	https://ccb.jhu.edu/software/centrifuge/
4	Taxonomic classification	Focus	An agile composition based approach using non-negative least squares (NNLS) to report the organisms present in metagenomic samples and profile their abundances.	https://peerj.com/articles/425/
4	Taxonomic classification	Gist	A statistical classifier for taxonomic inference for mRNA reads	https://github.com/rhetorica/gist

4	Taxonomic classification	graftm	A tool to identify and classify marker genes in short read datasets.	https://geronimp.github.io/graftM/
4	Taxonomic classification	GTDB-TK	A computationally efficient and able to classify thousands of draft genomes in parallel.	https://github.com/Ecogenomics/GTDBTk
4	Taxonomic classification	Kraken/ KRAKEN2	A taxonomic classification system using exact k-mer matches to achieve high accuracy and fast classification speeds.	https://ccb.jhu.edu/software/kraken2/
4	Taxonomic classification	MetaCV	A composition and phylogeny-based algorithm to classify very short metagenomic reads (75-100 bp) into specific taxonomic and functional groups.	https://sourceforge.net/projects/metacv/
4	Taxonomic classification	MetaPhyler	A novel taxonomic classifier for metagenomic shotgun reads, which uses phylogenetic marker genes as a taxonomic reference.	http://metaphyler.cbcb.umd.edu/
4	Taxonomic classification	PhymmBL	A new classification approach for metagenomics data which uses interpolated Markov models (IMMs) to taxonomically classify DNA sequences, c	https://www.cbcb.umd.edu/software/phymm/
4	Taxonomic classification	CCmetagen	CCMetagen processes sequence alignments to achieve read mappings. The pipeline is fast enough to use the whole NCBI nt collection as reference, facilitating the inclusion of understudied organisms in metagenome surveys.	https://github.com/vrmarcelino/CCMetagen/
5	Sequence assembly	AMOS/ MetAMOS	An open-source, modular assembly pipeline built upon AMOS and tailored specifically for metagenomic next-generation sequencing data	https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-s1-p25
5	Sequence assembly	BinSanity	A suite of scripts designed to cluster contigs generated from metagenomic assembly into putative genomes.	https://github.com/edgraham/BinSanity
5	Sequence assembly	Flye	A de novo assembler for single molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies.	https://github.com/fenderglass/Flye
5	Sequence assembly	GATB-minia-pipeline	A de novo assembly pipeline for Illumina data.	https://github.com/GATB/gatb-minia-pipeline

5	Sequence assembly	groopm	A metagenomics binning suite.	http://ecogenomics.github.io/GroopM/
5	Sequence assembly	IDBA-UD	Designed to utilize paired-end reads to assemble low-depth regions and use progressive depth on contigs to reduce errors in high-depth regions.	https://github.com/loneknightpy/idba https://pubmed.ncbi.nlm.nih.gov/22495754/
5	Sequence assembly	MaxBin/ MaxBin2	A software for binning assembled metagenomic sequences based.	https://toolshed.g2.bx.psu.edu/view/mbernt/maxbin2/cfd50144a871
5	Sequence assembly	MEGAHIT	An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.	https://github.com/voutcn/megahit
5	Sequence assembly	Meta-IDBA	Meta-IDBA algorithm for assembling reads in metagenomic data, which contain multiple genomes from different species	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3117360/
5	Sequence assembly	MetaBAT2	Clusters metagenomic contigs into different "bins", each of which should correspond to a putative genome.	https://kbase.us/applist/apps/metabat/run_metabat/release?gclid=Cj0KCQjwzbv7BRDIARIsAM-A6-2jVXdjGVpqsE23jl-nGvGJ81IBURBvM6dnevXoA06mQ42RPV_YqhkaAvevEALw_wcB
5	Sequence assembly	MetaCluster	Unsupervised binning method for metagenomic sequences.	https://github.com/mbanf/METACLUSTER
5	Sequence assembly	metaSPAdes	A versatile metagenomic assembler	http://spades.bioinf.spbau.ru/release3.11.1/manual.html https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5411777/
5	Sequence assembly	MetaVelvet	An extension of Velvet assembler to de novo metagenome assembly from short sequence reads	http://metavelvet.dna.bio.keio.ac.jp/ https://pubmed.ncbi.nlm.nih.gov/22821567/
5	Sequence assembly	MIRA	DNA sequence data assembler/mapper for whole genome and EST/RNASeq projects.	http://mira-assembler.sourceforge.net/docs/DefinitiveGuideToMIRA.html#sect_intro_whatismira
5	Sequence assembly	MyCC	MyCC is an automated binning tool that combines genomic signatures, marker genes and optional contig coverages within one or multiple samples, in order to identify the reconstructed genomic fragments.	https://sourceforge.net/projects/sb2nhri/files/MyCC/
5	Sequence assembly	S-GSOM	Binning sequences using very sparse labels within a metagenome.	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-215

5	Sequence assembly	SOAPdenovo2	A novel short-read assembly method that can build a de novo draft assembly for the human-sized genomes.	https://github.com/aquaskyline/SOAPdenovo2
5	Sequence assembly	SPADES - St. Petersburg genome assembler	An assembly toolkit containing various assembly pipelines.	https://cab.spbu.ru/software/spades/
5	Sequence assembly	Unicycler	An assembly pipeline for bacterial genomes.	https://github.com/rrwick/Unicycler
5	Sequence assembly	Velvet	A de novo genome assembler specially designed for short read sequencing technologies, such as Solexa or 454.	https://www.ebi.ac.uk/~zerbino/velvet/
6	Gene prediction and alignment	AMR++	A bioinformatics pipeline that interfaces with MEGARes to identify and quantify AMR gene accessions contained within a metagenomic sequence dataset.	https://academic.oup.com/nar/article/48/D1/D561/5624973
6	Gene prediction and alignment	BBMap	Splice-aware global aligner for DNA and RNA sequencing reads. It can align reads from all major platforms.	https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/
6	Gene prediction and alignment	BLAT	Accurate and 500 times faster than popular existing tools for mRNA/DNA alignments.	https://genome.cshlp.org/content/12/4/656
6	Gene prediction and alignment	BMGE - Block Mapping and Gathering with Entropy	Designed to select regions in a multiple sequence alignment that are suited for phylogenetic inference.	https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-210
6	Gene prediction and alignment	Bowtie/ Bowtie2	An ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.	http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#getting-started-with-bowtie-2-lambda-phage-example
6	Gene prediction and alignment	BWA	A software package for mapping low-divergent sequences against a large reference genome, such as the human genome.	http://bio-bwa.sourceforge.net/
6	Gene prediction and alignment	CD-HIT	A very widely used program for clustering and comparing protein or nucleotide sequences.	http://weizhongli-lab.org/cd-hit/
6	Gene prediction and alignment	DIAMOND	A sequence aligner for protein and translated DNA searches, designed for high performance analysis of big sequence data.	http://www.diamondsearch.org/index.php

6	Gene prediction and alignment	GlimmerMG	A system for finding genes in environmental shotgun DNA sequences.	http://www.cbcb.umd.edu/software/glimmer-mg/
6	Gene prediction and alignment	HMMER	Biosequence analysis using profile hidden Markov models.	http://hmmer.org/
6	Gene prediction and alignment	Infernal - INFERence of RNA ALignment	A useful tool for identifying RNAs in metagenomics data sets.	http://eddylab.org/infernal/
6	Gene prediction and alignment	IQ-TREE	Phylogenetic tree inference by maximum likelihood.	http://www.iqtree.org/
6	Gene prediction and alignment	MAFFT - Multiple Alignment with Fast Fourier Transform	A multiple sequence alignment program.	http://evomics.org/resources/software/bioinformatics-software/mafft/
6	Gene prediction and alignment	mauve	A system for constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion.	http://darlinglab.org/mauve/mauve.html
6	Gene prediction and alignment	MetaGene Annotator	A gene-finding program for prokaryote and phage.	http://metagene.nig.ac.jp/
6	Gene prediction and alignment	MetaGeneMark	Novel genomic sequences can be analyzed either by the self-training program GeneMarkS (sequences longer than 50 kb) or by GeneMark.hm.	http://exon.gatech.edu/meta_gmhmmp.cgi
6	Gene prediction and alignment	Minimap2	A general-purpose alignment program to map DNA or long mRNA sequences against a large reference database.	https://github.com/lh3/minimap2 https://academic.oup.com/bioinformatics/article/34/18/3094/4994778
6	Gene prediction and alignment	MinPath/ MinPath2	Minimal set of Pathways is for biological pathway reconstructions using protein family predictions.	https://omics.informatics.indiana.edu/MinPath/ http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000465
6	Gene prediction and alignment	NAST-iEr	Aligns a single raw nucleotide sequence against one or more NAST formatted sequences.	http://microbiomeutil.sourceforge.net/#A_NASTiEr

6	Gene prediction and alignment	PhyloSift	A suite of software tools to conduct phylogenetic analysis of genomes and metagenomes.	https://github.com/gjospin/PhyloSift
6	Gene prediction and alignment	PSORTm / PSORTb	For protein subcellular localization prediction (SCL).	https://www.psort.org/psortm/
6	Gene prediction and alignment	pyani	a Python package and standalone program for calculation of whole-genome similarity measures.	https://pyani.readthedocs.io/_/downloads/en/latest/pdf/
6	Gene prediction and alignment	TETRA	A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences.	https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-163
6	Gene prediction and alignment	tRNAscan-SE	The de facto tool for predicting tRNA genes in whole genomes.	http://trna.ucsc.edu/tRNAscan-SE/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768409/
7	Annotation prediction	BlastKOALA/ GhostKOALA	An automatic annotation server for genome and metagenome sequences, which perform KO (KEGG Orthology) assignments to characterize individual gene functions and reconstruct KEGG pathways.	https://www.sciencedirect.com/science/article/pii/S002228361500649X
7	Annotation prediction	dbCAN	A web server for automated Carbohydrate-active enzyme ANnotation.	http://bcb.unl.edu/dbCAN2/
7	Annotation prediction	eggNOG-mapper	A tool for fast functional annotation of novel sequences.	https://github.com/eggnogdb/eggnog-mapper
7	Annotation prediction	KAAS - KEGG Automatic Annotation Server	Provides functional annotation of genes by BLAST or GHOST comparisons against the manually curated KEGG GENES database.	https://www.genome.jp/kegg/kaas/
7	Annotation prediction	KofamKOALA	A web server to assign KEGG Orthologs (KOs) to protein sequences by homology search.	https://www.genome.jp/tools/kofamkoala/ https://academic.oup.com/bioinformatics/article/36/7/2251/5631907
7	Annotation prediction	PICRUSt/ PICRUSt2	A method to predict approximate functional potential of a community based on marker gene sequencing profiles.	https://github.com/picrust/picrust2 https://www.biorxiv.org/content/10.1101/672295v1.full

7	Annotation prediction	PROKKA	Annotation tool for bacterial, archaeal, and viral genomes.	http://www.metagenomics.wiki/tools/annotation/prokka
7	Annotation prediction	SUPER-FOCUS	A tool for metagenomics functional analysis, and it uses the SEED database.	https://github.com/metageni/SUPER-FOCUS
7	Annotation prediction	Tax4Fun2	An R-based tool for the rapid prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene marker gene sequences.	https://sourceforge.net/projects/tax4fun2/ https://www.biorxiv.org/content/10.1101/490037v1.full.pdf
8	Assembly Validation	CheckM	A set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes.	https://ecogenomics.github.io/CheckM/
8	Assembly Validation	CheckV	For assessing the quality of metagenome-assembled viral genomes.	https://www.biorxiv.org/content/10.1101/2020.05.06.081778v1
8	Assembly Validation	CompareM	A software toolkit which supports performing large-scale comparative genomic analyses. It provides statistics across sets of genomes (e.g., amino acid identity) and for individual genomes.	https://github.com/dparks1134/CompareM
8	Assembly Validation	Valet	Evaluating metagenomic assemblies.	https://github.com/marbl/VALET
9	Statistical analysis and visualisation	DADA2	Fast and accurate sample inference from amplicon data with single-nucleotide resolution.	https://benjjneb.github.io/dada2/index.html
9	Statistical analysis and visualisation	Krona	Allows hierarchical data to be explored with zooming, multi-layered pie charts.	https://github.com/marbl/Krona/wiki
9	Statistical analysis and visualisation	Metagenome Seq	Designed to determine features (be it Operational Taxonomic Unit (OTU), species, etc.) that are differentially abundant between two or more groups.	https://www.bioconductor.org/packages/release/bioc/html/metagenomeSeq.html
9	Statistical analysis and visualisation	MetaPath	Identify differentially abundant pathways in metagenomic data-sets.	https://www.cbcb.umd.edu/software/metapath
9	Statistical analysis and visualisation	Phyloseq	A set of classes and tools to facilitate the import, storage, analysis, and graphical display of microbiome census data.	https://www.bioconductor.org/packages/release/bioc/html/phyloseq.html
10	Databases	CAzy - Carbohydrate-A	Describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes that	http://www.cazy.org/

		ctive enZYmes Database	degrade, modify, or create glycosidic bonds.	
10	Databases	COG Clusters of Orthologous Groups of proteins	A developed system for delineation of Clusters of Orthologous Groups of proteins (COGs) from the sequenced genomes of prokaryotes and unicellular eukaryotes and the construction of clusters of predicted orthologs.	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC222959/
10	Databases	Cyanorak	Cyanorak Information system is a bioinformatics tool dedicated to the curation, comparison and visualization of genomes of strains belonging to the subsection I, cluster 5, a deeply branching group within the Cyanobacteria phylum.	http://application.sb-roscoff.fr/cyanorak/;jsessionid=80A8351DD8C9E9DCFFC029AAE3BDF83A?execution=e1s1
10	Databases	EBI	European Bioinformatics Institute.	https://www.ebi.ac.uk/
10	Databases	eggNOG	A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.	http://eggnog5.embl.de/#/app/home
10	Databases	FunGuild	A python-based tool that can be used to taxonomically parse fungal OTUs by ecological guilds independent of sequencing platforms or analysis pipelines.	http://www.funguild.org/
10	Databases	Greengenes	16S rRNA gene database or experimental datasets.	https://greengenes.secondgenome.com/
10	Databases	GTDB	Genome taxonomy database.	https://gtdb.ecogenomic.org/
10	Databases	InterPro	Functional analysis of proteins by classifying them into families and predicting domains and important sites.	https://www.ebi.ac.uk/interpro/
10	Databases	KEGG: Kyoto Encyclopedia of Genes and Genomes KEGG	KEGG is a database resource for understanding high-level functions and utilities of the biological system	https://www.genome.jp/kegg/
10	Databases	KOG eukaryotic orthologous groups (KOGs)	A eukaryote-specific version of the Clusters of Orthologous Groups (COG) tool for identifying ortholog and paralog protein	https://mycocosm.jgi.doe.gov/Tutorial/tutorial/kog.html https://www.hsls.pitt.edu/obrc/index.php?page=URL1144075392

10	Databases	MAR	Marine databases; MarRef, MarDB and MarCat, which are publicly available resources that promote marine research and innovation.	https://mmp.sfb.uit.no/databases/ https://academic.oup.com/nar/article/46/D1/D692/4584637
10	Databases	MEROPS	An information resource for peptidases (also termed proteases, proteinases and proteolytic enzymes) and the proteins that inhibit them.	https://www.ebi.ac.uk/merops/ https://academic.oup.com/nar/article/46/D1/D624/4626772
10	Databases	MetaCyc	A curated database of experimentally elucidated metabolic pathways from all domains of life.	https://metacyc.org/
10	Databases	NCBI	National Center for Biotechnology Information.	www.ncbi.nlm.nih.gov
10	Databases	PANTHER - Protein ANalysis THrough Evolutionary Relationships)	Designed to classify proteins (and their genes) in order to facilitate high-throughput analysis.	http://www.pantherdb.org/data/
10	Databases	Pfam	A large collection of protein families.	https://pfam.xfam.org/
10	Databases	PR2	A reference database of carefully annotated 18S rRNA sequences using eight unique taxonomic fields.	https://pr2-database.org/
10	Databases	RDP	Provides the research community with aligned and annotated rRNA gene sequence data.	http://rdp.cme.msu.edu/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965039/
10	Databases	Rfam	A collection of RNA families, each represented by multiple sequence alignments.	https://rfam.xfam.org/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4383904/
10	Databases	SEED	To provide consistent and accurate genome annotations across thousands of genomes and as a platform for discovering and developing de novo annotations.	https://pubseed.theseed.org/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965101/
10	Databases	Silva	A comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (Bacteria, Archaea and Eukarya).	https://www.arb-silva.de/

10	Databases	TARA Oceans	Diversity, evolution and ecology of marine plankton.	https://www.ebi.ac.uk/services/tara-oceans-data http://www.taraoceans-dataportal.org/top/;jsessionid=07217630362165E3CD27AA73D839945D?execution=e1s1
10	Databases	TCDB	A comprehensive IUBMB approved classification system for membrane transport proteins known as the Transporter Classification (TC) system.	http://www.tcdb.org/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1334385/
10	Databases	TIGRFAM	A resource consisting of curated multiple sequence alignments, Hidden Markov Models (HMMs) for protein sequence classification, and associated information designed to support automated annotation of (mostly prokaryotic) proteins.	http://tigrfams.jcvi.org/cgi-bin/index.cgi
11	Other	Anvi'o	An open-source, community-driven analysis and visualization platform for microbial ‘omics.	http://merenlab.org/software/anvio/
11	Other	Calypso	An easy-to-use online software, allowing non-expert users to mine, interpret and compare taxonomic information from metagenomic or 16S rDNA datasets.	http://cgenome.net/wiki/index.php/Calypso
11	Other	CLC Genomics Workbench	A bioinformatics software solution that allows for comprehensive analysis of your NGS data, including de novo assembly of whole genomes and transcriptomes, resequencing analysis.	https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-clc-genomics-workbench/
11	Other	conda	An open source package management system and environment management system that runs on Windows, macOS and Linux.	https://docs.conda.io/en/latest/
11	Other	Galaxy Australia	Galaxy is a web-based analysis and workflow platform.	https://usegalaxy.org.au/
11	Other	gromacs	A versatile package to perform molecular dynamics.	http://www.gromacs.org/
11	Other	IMG/M	A platform to support the annotation, analysis and distribution of microbial genome and microbiome datasets.	https://img.jgi.doe.gov/
11	Other	Jupyter Notebook	A open-source web application that allows you to create and share documents that contain live code,	https://jupyter.org/

11	Other	MEGAN - MEtaGenome ANalyzer	A comprehensive toolbox for interactively analyzing microbiome data.	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/megan6/
11	Other	MetaORFA - Metagenomic ORFome Assembly	Metagenomic assembly.	http://allie.dbcls.jp/pair/MetaORFA;Metagenomic+ORFome+Assembly.html
11	Other	MetaWRAP	An easy-to-use metagenomic wrapper suite that accomplishes the core tasks of metagenomic analysis from start to finish.	https://github.com/bxlab/metaWRAP https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0541-1
11	Other	MG-RAST	An automatic phylogenetic and functional analysis of metagenomes.	https://www.mg-rast.org/
11	Other	MGnify	An analysis, archiving and browsing of metagenomic and metatranscriptomic data.	https://www.ebi.ac.uk/metagenomics/
11	Other	MOCAT/ MOCAT2	A package for analyzing metagenomics datasets.	https://mocat.embl.de/
11	Other	Mothur	An open-source, expandable software to fill the bioinformatics needs of the microbial ecology community.	https://www.mothur.org/
11	Other	Nextflow	A scalable and reproducible scientific workflow using software containers.	https://www.nextflow.io/
11	Other	OTUreporter	A modular automated pipeline for the analysis and report of amplicon data.	https://bitbucket.org/xvazquezc/otureporter/wiki/Home
11	Other	Perl	A general purpose language for getting things done.	https://www.perl.com/about/
11	Other	Python	Programming language	https://www.python.org/
11	Other	QIIME2.0	Performing microbiome analysis from raw DNA sequencing data.	https://qiime2.org/
11	Other	R/R Studio	A development environment for R and Python, with a console, syntax-highlighting editor.	https://rstudio.com/
11	Other	RocksDB	A persistent key-value store for flash and RAM storage	https://github.com/facebook/rocksdb

11	Other	singularity	Singularity containers can be used to package entire scientific workflows,	https://singularity.lbl.gov/
11	Other	SOAP - Short Oligonucleotide Analysis Package	A suite of bioinformatics software tools from the BGI Bioinformatics department enabling the assembly, alignment, and analysis of next generation DNA sequencing data.	http://manpages.ubuntu.com/manpages/cosmic/man1/soap.1.html
11	Other	SqueezeMeta	A fully automatic pipeline for metagenomics/metatranscriptomics, covering all steps of the analysis.	https://github.com/jtamames/SqueezeMeta https://www.frontiersin.org/articles/10.3389/fmicb.2018.03349/full#h2
11	Other	VAMPS	A collection of tools for researchers to visualize and analyze data for microbial population structures and distributions.	https://vamps2.mbl.edu/

A complete list of tools with more details is available here.

Appendix 2

Survey[89] questions posed to the Microbiome Research Community

1. How would you describe your level of experience with metagenomics or microbiome analysis?

- ❏ Very experienced
- ❏ Some experience
- ❏ Beginner
- ❏ Interested but no direct experience
- ❏ Other:

2. Which part(s) of the analysis process do you / group members perform, or envisage performing in the next 5 years?

- ❏ Targeted amplicon sequencing
- ❏ Random shotgun sequencing
- ❏ Taxonomic profiling
- ❏ Functional profiling
- ❏ Generating metagenome-assembled genomes (MAGs)
- ❏ Phylogenetic analysis
- ❏ Statistical analyses
- ❏ Novel gene discovery
- ❏ Other:

3. With respect to metagenomics or microbial analyses, which host/habitat/environment have you sampled from (or work on), or will sample from in the next 5 years (choose all that apply)?

- ❏ Human host-associated samples (e.g. fecal sample from human)
- ❏ Non-human host-associated samples (e.g. fecal sample from a koala, leaf-associated sample from a plant)
- ❏ Marine or freshwater biome samples (e.g. river, rainwater, ocean, estuarine, tap water, etc)
- ❏ Terrestrial environmental biome samples (e.g. desert, forest, mangrove, cropland, urban, etc)
- ❏ Other:

4. Which of the following reference databases do you use (choose all that apply)? **NB. this list is non-exhaustive so please note preferences not listed in 'other'

- ❏ COG/KOG
- ❏ EBI
- ❏ eggNOG

[89] Metagenome and Microbiome Poll/Survey

- ❏ Greengenes
- ❏ KEGG
- ❏ Mockrobiota
- ❏ NCBI
- ❏ PFAM
- ❏ RDP: Ribosomal Database Project
- ❏ SEED
- ❏ Silva
- ❏ TIGRFAM
- ❏ Custom-made database
- ❏ Other:

5. Which (if any) tools / software / pipelines / programs / platforms do you or group members use (choose all that apply)? Please only indicate those you'd currently recommend for use . **NB. this list is non-exhaustive so please note preferences not listed in 'other'

- ❏ AMOS (A Modular Open-Source Assembler)
- ❏ ANVI'O
- ❏ BWA
- ❏ Bowtie or Bowtie2
- ❏ CLC Genomics Workbench
- ❏ CD-HIT
- ❏ BLAST+
- ❏ BLAT
- ❏ BlastKOALA and/or GhostKOALA (KOALA: KEGG Orthology And Links Annotation)
- ❏ DiScRIBinATE
- ❏ FastQC
- ❏ Fastx-Toolkit
- ❏ FragGeneScan
- ❏ Galaxy Australia
- ❏ GlimmerMG
- ❏ Genometa
- ❏ HMMER
- ❏ IDBA-UD
- ❏ IMG
- ❏ Jupyter Notebook
- ❏ Kaggle
- ❏ KAAS (KEGG Automatic Annotation Server)
- ❏ LotuS and sdm (less OTU scripts and simple demultiplexer)
- ❏ MaxBin
- ❏ MED
- ❏ MEGAN
- ❏ MetaGeneAnnotater (MGA)/ Metagene

- ❏ MetagenomeSeq
- ❏ Meta-IDBA
- ❏ MetaORFA
- ❏ MetaPath
- ❏ META-PIPE
- ❏ METASPADES
- ❏ MetaVelvet
- ❏ Meta-QC-Chain
- ❏ MetaCluster
- ❏ MetaPhyler
- ❏ MGnify (EBI Metagenomics)
- ❏ MG-RAST
- ❏ MinPath
- ❏ MIRA
- ❏ MOCAT
- ❏ MOTHUR
- ❏ Parallel-meta
- ❏ PEAR
- ❏ PICRUSt or PICRUSt2
- ❏ ProViDE
- ❏ PROKKA
- ❏ PCAHIER
- ❏ Phyloseq
- ❏ PhymmBL
- ❏ Python
- ❏ QIIME or QIIME2
- ❏ R / R Studio
- ❏ RAMMCAP
- ❏ Ray Meta
- ❏ SPARCC
- ❏ ShotgunFunctionalizeR
- ❏ SORT-Items
- ❏ SOAP
- ❏ SPADES
- ❏ S-GSOM
- ❏ SOrt-ITEMS
- ❏ TETRA
- ❏ TACAO
- ❏ USEARCH
- ❏ VSEARCH
- ❏ Velvet
- ❏ VAMPS
- ❏ Custom tool developed in our group or by collaborator

- ❏ Other:

6. Are there tools / software / pipelines / programs / platforms you'd like to use but that aren't suitable for your study taxon/taxa? If so, what are they and why aren't they suitable?

7. Are there tools / software / pipelines / programs / platforms you'd like to use but can't because of technical limitations (e.g. installation, compute requirements, dataset access requirements)? If so, what are the tools and what are the roadblocks you've encountered? What is your workaround and why is it inadequate?

8. Do you require custom or proprietary tools / software for your metagenomics approach? If so, what are they?

9. What sequencing platform/s are you currently using to generate data (choose all that apply)?

- ❏ Illumina
- ❏ PacBio
- ❏ 10 X
- ❏ Nanopore
- ❏ Ion Torrent
- ❏ Other:

10. Do you make use of existing datasets from the same taxon or closely related taxa (choose all that apply)?

- ❏ Yes, public datasets from the same taxon
- ❏ Yes, private datasets from the same taxon (from my previous work or that of collaborators)
- ❏ Yes, public datasets from closely related taxa
- ❏ Yes, private datasets from closely related taxa (from my previous work or that of collaborators)
- ❏ No, because no relevant data exists from my taxon or a closely-related taxon
- ❏ No - some data exists but it's too low quality for this purpose
- ❏ No - some data exists but it's too difficult to integrate because of poor/outdated format or metadata
- ❏ No - some data exists but it's too difficult to integrate because of a lack of suitable tools/pipelines
- ❏ No - some private data exists but I can't access it
- ❏ Other:

11. Do you use a data management tool/framework within your metagenomics project(s)? If so, what?

12. How do you share data within your group and with collaborators? Where are your collaborators based? What difficulties have you encountered?

13. Do you make your metagenomic datasets publicly available? If so, where? Have you encountered any difficulties in doing so?

14. If you don't make your metagenomic datasets publicly available, why not?

- ❏ Commercial confidence issues
- ❏ I don't see a benefit in sharing my metagenomic datasets publicly available
- ❏ I don't know how to make my metagenomic datasets publicly available
- ❏ Other:

15. What kind of compute infrastructure setup do you use for metagenomics (choose all that apply)?

- ❏ Local desktop/PC
- ❏ High-performance computing at my institution
- ❏ High-performance computing at a collaborator's institution
- ❏ High-performance computing within my research group
- ❏ High-performance computing within my Department
- ❏ National or state high-performance computing infrastructure (e.g. NCI, Pawsey, QCIF/QRIScloud) NeCTAR cloud instance Commercial cloud (e.g. Amazon Web Services, Microsoft Azure, Google Cloud) Galaxy

16. Do you have access to the expertise you need to build and maintain this compute infrastructure (e.g. installing and updating software)?

17. Is your current compute infrastructure sufficient for your current needs? If not, why not?

18. Will this compute infrastructure setup be sufficient for your needs in 2 years' time?

19. Will this compute infrastructure setup be sufficient for your needs in 5 years' time?

20. Would you / group members use a shared compute infrastructure platform to perform metagenomics?

21. How important are these general factors to you in a shared metagenomics platform?

- ❏ Following best practice in tools, formats and metadata; compliant with requirements of international data repositories
- ❏ Free (subsidised) to researchers no matter the scale of analysis
- ❏ Easy to access from anywhere
- ❏ Easy to self-manage access and permissions for collaborators
- ❏ Easy to upload/download data
- ❏ Security of data and analysis
- ❏ Long-term support for and sustainability of the platform

22. How important are these data-related factors to you in a shared metagenomics platform?

- ❏ Smart metadata handling (e.g. assistance with metadata formats, transfer of metadata through pipeline, controlled vocabulary lookup)
- ❏ Ability to submit datasets to international repositories from the platform
- ❏ Ability to download datasets from international repositories within the platform
- ❏ Ability to transfer data easily to/from storage

26. How important are these training-related factors to you in a shared metagenomics platform?

- ❏ Good documentation on how to use the platform
- ❏ Good documentation on how to use the tools/pipelines
- ❏ Access to in-person training on how to use the platform
- ❏ Access to in-person training on how to use the tools/pipelines
- ❏ Discussion forum to share expertise with other users

23. How important are these tool/pipeline-related factors to you in a shared metagenomics platform?

- ❏ Access to our preferred tools/pipelines
- ❏ Access to a choice of tools/pipelines
- ❏ Quick installation of other tools/pipelines upon request
- ❏ Assistance available in implementing pipelines

24. What are the top 1-5 tools/pipelines you would absolutely require in a shared metagenomics platform?

25. How important are these compute-related factors to you in a shared metagenomics platform?

- ❏ Ability to scale up/down resources used as needed
- ❏ No need to understand or control the compute backend
- ❏ Compatibility with external analysis environments (e.g. Amazon, Cyverse)

27. Are there any other factors you consider crucial in a shared metagenomics platform? If so, what?

Document Control

VERSION	DATE	AUTHOR(S)	DESCRIPTION
V1.0	30/11/2020	Tiffanie Nelson, Jeff Christiansen	A preliminary document detailing the outline of the roadmap draft including the software list obtained from researchers.